Nama : Ilham Wahyu Adli

Nim : 1301173380

Kelas : IF Gab-02

Pada tugas *semantic vector* ini saya memilih topik "*Staycation*" dan "Kebakaran Hutan". Artikel yang digunakan adalah beberapa kalimat dari artikel yang didapat dalam beberapa portal berita. Portal berita yang dipilih antara lain adalah: tempo, detik dan kompas. Untuk pemilihan artikelnya saya menggunakan mesin pencarian Google dengan *keyword* topik yang telah ditentukan. Setelah mendapatkan satu artikel dari salah satu portal berita, saya melihat rekomendasi berita lainnya dengan topik yang sama dari halaman tersebut sehingga ada kemungkinan jika artikel tersebut mirip.

Ukuran matriks yang dipilih

Dikarenakan penggunaan artikel hanya berupa beberapa kalimat dari suatu artikel dan penggunaan *library* Sastrawi untuk melakukan *preprocessing* artikel seperti penghapusan *stopword* dan melakukan *stemming* kata. Total kata unik yang didapatkan setelah melakukan *preprocessing* adalah sebanyak 381 kata. Oleh karena itu saya memutuskan untuk menggunakan keseluruhan kata unik tersebut. Sehingga ukuran matriks yang dibuat pada pembuatan tugas 3 ini adalah:

1. Matriks TF-IDF = 20 baris x 381 kolom
2. Matriks Co-Occurrence Term Context = 381 baris x 381 kolom
3. Matriks PPMI = 381 baris x 381 kolom

Hasil Implementasi

Implementasi pertama dalam tugas ini adalah menghitung persen dari elemen matriks yang bernilai tidak sama dengan 0. Persentase matriks TF-IDF memiliki nilai lebih besar yaitu 7,76% dibandingkan matriks PPMI yang memiliki nilai 3,8%. Hal ini dikarenakan ukuran matriks yang berbeda sehingga pada matriks PPMI memiliki kemungkinan nilai kosong lebih banyak.

```
Start Nomor 1
7.76% Tidak bernilai Nol
3.8% Tidak bernilai Nol
End Nomor 1
```

Gambar 1 Perbandingan Matriks TF-IDF (7,76%) dengan Matriks PPMI (3,8%)

Kemudian yang selanjutnya dilakukan adalah membandingan dokumen pada matriks TF-IDF dengan perhitungan *cosine similarity* hasil yang didapat adalah seperti dibawah ini:

```
Start Nomor 2
Persamaan antara dokumen 5 dan 5 (Dokumen Sama):  1.0000000000000002
Persamaan antara dokumen 12 dan 20 (Dokumen dengan Topik Sama):  0.23086674460441065
Persamaan antara dokumen 3 dan 13  (Dokumen dengan Topik Berbeda):  6.005131832483395e-08
End Nomor 2
```

Gambar 2 Implementasi Soal 2

Dapat dilihat ketika dokumen tersebut identik yaitu dokumen 5 dan 5 mendapatkan nilai 1 dimana dokumen tersebut sama. Ketika menggunakan dokumen 12 dan 20 yang membahas

topik *Staycation* nilai *cosine similarity*-nya menjadi 0.23 hal ini membuktikan adanya sedikit kesamaan dari dokumen 12 dan 20. Dan ketika menggunakan dokumen dengan topik yang berbeda yaitu dokumen 3 dan 13 nilainya sangat jauh dari angka 1, hal ini membuktikan bahwa penggunaan kata dalam dokumen 3 dan 13 sangat berbeda dikarenakan topik yang jauh berbeda pula.

Implementasi akhir dari tugas ini adalah membandingkan suatu kata dengan kata lainnya. Yang diharapkan adanya kemiripan dari 2 kata tersebut. Dengan menggunakan ketiga matriks yang telah dibangun dan kata pertama yang dipilih adalah "*Staycation*" dan kata kedua adalah "Liburan". Hasil yang diperoleh adalah seperti dibawah ini:

```
Perbandingan Kata (Staycation dan Liburan)
Implementasi Matriks TF-IDF:  0.17719882804427356
Implementasi Matriks Term Context:  0.7617170605358935
Implementasi Matriks PPMI:  1.5031001874813201
```

Gambar 3 Perbandingan Pertama Antar kata Dengan Matriks yang Berbeda

Ketika menggunakan matriks TF-IDF untuk mencari kemiripan kata hasil yang didapatkan sangat kecil, sedangkan ketika menggunakan Term Context hasilnya mendekati 1. Hal ini dikarenakan ketika menggunakan implementasi TF-IDF kata tersebut hanya berkemungkinan muncul pada 1 dokumen saja. Sehingga 10 elemen matriks lainnya memiliki nilai 0. Sedangkan ketika menggunakan Term Context dimana menggunakan implementasi kedekatan kata, disini penulis menggunakan implementasi yang dilakukan oleh Daniel Jurafsky ketika membangun Term Context yaitu mencari 4 kata dari sebelah kiri dari posisi kata yang ingin dicari dan juga 4 kata dari sisi kanannya. Sehingga kemunculan dari kata *staycation* dan Liburan menjadi sangat tinggi, karena *staycation* dan liburan merupakan suatu kesatuan yaitu *staycation* merupakan jenis aktifitasnya dan liburan merupakan konteksnya. Dapat dibuktikan juga ketika memasukan 2 kata yang berbeda topiknya seperti "*Staycation*" dan "Kebakaran". Hasil yang didapat ketika menggunakan 2 kata tersebut adalah:

```
Perbandingan Kata (Staycation dan Kebakaran)
Implementasi Matriks TF-IDF:  0.0
Implementasi Matriks Term Context:  0.05714112240750387
Implementasi Matriks PPMI:  0
```

Gambar 4 Perbandingan Kedua Antar Kata Dengan Matriks yang Berbeda

Dapat dilihat dari hasil tersebut kata "*Staycation*" dan "Kebakaran" tidak memiliki kemiripan sehingga nilai dari ketika matriks tersebut sangat jauh lebih kecil dibandingkan ketika menggunakan kata yang saling berkaitan.